Do Our Psychological Laws Apply Only to College Students?

Stroebe, Wolfgang; Gadenne, Volker; Nijstad, Bernard A.

Published in:
Basic and Applied Social Psychology

DOI:
10.1080/01973533.2018.1513362

Publication date:
2018

Basic and Applied Social Psychology

ISSN: 0197-3533 (Print) 1532-4834 (Online) Journal homepage: https://www.tandfonline.com/loi/hbas20

Do Our Psychological Laws Apply Only to College Students?: External Validity Revisited

Wolfgang Stroebe, Volker Gadenne & Bernard A. Nijstad

To cite this article: Wolfgang Stroebe, Volker Gadenne & Bernard A. Nijstad (2018) Do Our Psychological Laws Apply Only to College Students?: External Validity Revisited, Basic and Applied Social Psychology, 40:6, 384-395, DOI: 10.1080/01973533.2018.1513362

To link to this article: https://doi.org/10.1080/01973533.2018.1513362

Published online: 15 Oct 2018.

BASIC AND APPLIED SOCIAL PSYCHOLOGY
2018, VOL. 40, NO. 6, 384–395
https://doi.org/10.1080/01973533.2018.1513362

Routledge
Taylor & Francis Group

Do Our Psychological Laws Apply Only to College Students?: External Validity Revisited

Wolfgang Stroebe[a], Volker Gadenne[b], and Bernard A. Nijstad[a]

[a]University of Groningen; [b]University of Linz

ABSTRACT

That most psychological research is conducted with students led to concerns that psychological laws apply only to this population. These fears are based on Campbell and Stanley's concept of external validity that specifies the extent to which research findings can be generalized. This concept is based on an inductivist philosophy. As philosophers of science have argued since Hume, one cannot derive general laws from singular observations. Instead, one develops theories and uses empirical studies to test these theories. This solves the problem of generalization because the domain of applicability is specified by the theory. Reports that studies result in different findings when conducted in different cultures are unproblematic as long as these differences can be explained with psychological theories.

There seems to be a growing concern in psychology about the lack of diversity in the subject populations on which we test our theories. For example, Henrich, Heine, and Norenzayan (2010) criticized in a widely cited article that "behavioral scientists routinely publish broad claims about human psychology and behavior. based on samples drawn entirely from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies" (p. 61). Arnett (2008) likewise noted that most conclusions from psychological research are based on American samples and that conclusions that "apply to all human beings" have to be "based on studies of diverse sectors of the human population" (p. 602). It has even been suggested that this lack of diversity may be at the root of our so-called replication crisis.[1] Thus, Simons, Yuichi, and Lindsay (2017) argued that replication failures could often be due to researchers failing to sample from the same subject population that had been employed in the original study. Along similar lines, Kitayama (2017), as the incoming editor of *Journal of Personality and Social Psychology* (Attitudes and Social Cognition section), stated in his editorial that "if the same effects occur in a sample that is very different from the original one, this will constitute a 'big plus' that could bring the paper above the threshold for publication in JPSP:ASC" (p. 359).

This renewed interest for the (lack of) diversity in subject populations stems partly from cross-cultural research showing that psychological findings often do not generalize across cultures (e.g., Henrich et al., 2010). It also partly stems from a sincere desire to improve scientific practices within psychology, along with more attention for statistical power, replication, and transparency (e.g., Funder et al., 2014). For example, referring to the replication crisis, Kitayama (2017) noted, "Ultimately, I believe that this intentional expansion of the subject base—not only in size but also in diversity—in our science is the best step toward addressing the challenges we face today" (p. 359). Despite these good intentions, in this article we argue that the assumption that the populations on which we test our theories need to be representative of the population to which these theories are applied is wrong. As a consequence, "blindly" (i.e., without theory) replicating findings among different populations is likely a waste of resources.

The notion of external validity

The erroneous belief that, in scientific research, subject populations need to be "representative" has a long tradition in scientific discourse but received methodological justification through the concept of *external validity*. This concept was developed originally by

CONTACT Wolfgang Stroebe wolfgang.stroebe@gmail.com Department of Social and Organizational Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.

Campbell in an article in *Psychological Bulletin* in 1957 but became popular through a little booklet entitled "Experimental and Quasi-Experimental Designs for Research" (Campbell & Stanley, 1966). This little booklet—the reprint of a chapter that Campbell and Stanley wrote for the *Handbook of Research on Teaching* (Gage, 1963)—was one of the most influential methodological contributions to psychological research ever published. It constituted compulsory reading for generations of graduate students. Partly due to its accessible writing, it probably contributed more to the improvement of research in the social sciences than any other methodological publication.

Campbell and Stanley (1966) popularized the distinction between two types of validity, namely, internal validity and external validity. They stated that

> internal validity is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? External validity asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized? (Campbell & Stanley, 1966, p. 5)

The fact that the booklet is a reprint of a chapter written for a handbook on education research is important because their discussion of external validity has to be understood in the context of applied research, where, as we argue next, external validity is more relevant than for theory testing. Campbell and Stanley (1966) did not even mention problems involved in basing general laws[2] on studies of college students, an issue that later became the main focus of this discussion. Even though the unease with the lack of diversity in the research that participants studied in psychological research predates the Campbell and Stanley (1963/1966) treatise, the need for external validity has often been used as justification for the critique that psychology, and particularly social psychology, had become the science of the American college student and that psychological laws might not be valid outside this limited population (e.g., Arnett, 2008; Henrich et al., 2010; Simons et al., 2017).

The problem with the concept of external validity is that it is based on *inductivist principles*. According to the first inductivist principle, general statements (laws, theories) are generated by deriving them from observation statements, usually by generalizing the results of a limited set of observations. The second inductive principle states that a general statement formed this way can be considered as proven, if the number of observations is sufficiently great, a variety of cases were observed, and no observation statement contradicted the general statement. Under these conditions, the inference from the observation statement to the general statement is regarded as valid or justified.

The first inductive principle is unproblematic, because it is concerned only with the question of how to generate theories, not with their truth or justification. Induction may be used as a heuristic for theory building. The usefulness of induction for that purpose is, however, controversial: Some consider it as very useful, but others point out that explanatory theories, which refer to unobservable structures and processes, can hardly be created by generalizing observable states of affairs. But we need not discuss this here, as the problem of induction is not mainly associated with the formation of theories.

It is the second principle that was famously criticized by Hume (1748; see also Popper, 1959). An inference from singular statements to a general statement cannot be founded on deductive logic. Moreover, the attempt to justify induction by arguing that induction has proved successful in the past is circular. An inference from past success to future success is itself an inductive inference. Popper also demonstrated that the same problem arises if the inductive inference or its conclusion is only claimed as probable. As a result of this criticism, scientific theories can never be proven to be true or probable. A theory can get confirmation (or support) from observations, however, only of observations that are the result of tests that could have refuted the theory at stake, not from observations that gave rise to its very formulation.[3]

Furthermore, external validity and the associated idea of generalizability are concerned not only with the formation of general statements but also with their *truth or justification*. The concept of validity clearly indicates that some generalizations are regarded as valid, whereas others are not. But how could a generalization be valid? If Hume's argument is sound, one can simply *never* generalize research findings in a sample to a population, or in a specific population to other populations. In other words, when a researcher finds a specific relation between two variables, it is never possible to logically prove that this relation will also hold in a different population (or when using different operationalizations of the variables involved).

That the concept of external validity is based on a faulty inductivist philosophy of science has been recognized and severely criticized soon after the Campbell and Stanley (1966) publication (Calder, Phillips, & Thybout, 1982; Gadenne, 1976; Kruglanski & Kroy, 1973; Mook, 1983). The remarkable aspect of these publications is not only their strong reservations

about the construct of external validity but also the fact that these authors developed their arguments independent of one another. For example, Mook (1983) flatly argued that the statement that "the selection of designs strong in both types of validity is obviously our ideal" (Campbell & Stanley, 1966, p. 5) is wrong. Instead, he argued, when testing theories, one is not even interested in generalizability: It is not important whether specific findings are also found among other populations; what is important is drawing valid conclusions about a theory given the empirical evidence. Along similar lines, Calder et al. wrote,

> It seems self-evident to many researchers, for instance, that having a random sample from some larger population is a better test than employing a convenience (e.g., student) sample. Yet much of this superficial plausibility disappears on closer inspection. The reason is that theories are stated at a universal level. As long as a sample is relevant to the universe of the theory, it constitutes a test of that theory. (p. 241)

Even Cook and Campbell (1979) expressed some doubts about the importance of external validity for theory-testing research. In their follow-up of Campbell and Stanley (1966) treaty, they wrote, "The priority among validity types varies with the kind of research being conducted. … Few theories specify crucial target settings, populations, or times across which generalization is desired. Consequently, external validity is of relatively little importance" (p. 83). However, even though they agreed that it is less important in theory-testing research, they could not quite abandon they idea that it is at least of some importance: "For investigators with theoretical interests our estimate is that the types of validity, in order of importance, are probably internal, construct, statistical, and external validity" (p. 83).

The fact that this confusion on an issue that is central to psychological research is still alive justifies a renewed discussion of the question of generalizability and external validity. Such a discussion seems particularly timely, because recent findings from cross-cultural psychology have created new doubts about the general validity of psychological theories (Henrich et al., 2010).

Psychology as the study of college students: Subject selection bias in psychological research

Complaints about the lack of diversity in the populations of research participants used in psychological experiments have a long tradition. Probably the first psychologist to make that point was McNemar (1946), who, in an article on opinion-attitude methodology, argued that

> the existing science of human behavior is largely the science of the behavior of sophomores. Too much research effort is expended on college students with subsequent waste of journal space devoted to speculation concerning whether the findings hold for mankind in general. (p. 333)

In retrospect, McNemar's complaint seems premature, because even in 1949 the proportion of studies with undergraduate participants published in the *Journal of Abnormal and Social Psychology* was only 20%. It had doubled to 49% by 1959 (Christie, 1965) and rose to 73% by 1962 to 1964 (Smart, 1966). Because the majority of these student subjects were male, Smart (1966) concluded that "the male college student has become the 'white rat' of human experimentation" (p. 115). Smart could have added "American" to this description, because in an analysis of articles published from 2003 to 2007 in six journals of the American Psychological Association, Arnett (2008) found that approximately 80% of research participants were American. Although there was little change in this situation during the last century (Wintre, North, & Sugar, 2001), the advance of Amazon.com's Mechanical Turk (MTurk) and other online sources (e.g., Prolific Academic) has made other easily accessible subject pools available (Paolacci & Chandler, 2014). Although the use of MTurk will increase the diversity of research participants taking part in a study, individuals who earn money by serving as participants in psychological studies are likely to represent an atypical slice of humanity.

The unease with basing psychological laws on studies of the small minority of individuals who attend college has been justified with arguments that these students are relatively exceptional (and homogeneous) in many characteristics that are associated with psychological factors, such as age, social class, and learning ability (Smart, 1966). The most elaborated critique was published by Sears (1986) in the *Journal of Personality and Social Psychology*. He argued that

> research on the full life span suggests that, compared with older adults, college students are likely to have less crystallized attitudes, less formulated senses of self, stronger cognitive skills, stronger tendencies to comply with authority, and more unstable peer group relationships. The laboratory setting is likely to exaggerate all these differences. These peculiarities of social psychology's predominant data base may have contributed to central elements of its portrait of human nature. (p. 515)

A few decades later, Henrich et al. (2010) presented data from cross-cultural research that demonstrated

that many findings based on Western participants cannot be replicated with non-Westerners. Even such basic perceptual effects as the Müller-Lyer visual illusion were different for Western as compared to non-Western participants.

To summarize, until very recently the overwhelming proportion of psychological studies was based on college student participants, most of them studying at U.S. universities. This situation has changed with the rise of crowdsourced subject pools such as Amazon's MTurk. But although the recent surge in use of MTurk samples has increased diversity, it arguably did not improve representativeness. Finally, research suggests that many psychological findings may not generalize to other cultures. This state of affairs has led some to propose that psychology needs to diversify its subject populations. We argue, however, that this is a misconception, based on erroneous inductivist logic: Diversification of subject populations does not make experimental findings more externally valid.

Inductivist strategies to derive general laws from experimental observations

In discussing factors jeopardizing external validity, Campbell and Stanley (1966) were aware of the fact that this concept was problematic. They warned that a caveat was in order that introduced

> some painful problems in the science of induction. The problems are painful because of a recurrent reluctance to accept Hume's truism that induction or generalization is never fully justified logically. Whereas the problems of internal validity are solvable within the limits of the logic of probability statistics, the problems of external validity are not logically solvable in any neat, conclusive way. (p. 17)

Even though the ideas developed in Campbell and Stanley (1966) were further expanded by Cook and Campbell (1979) and Shadish, Cook, and Campbell (2002), their refinement of the concept of validity failed to amend the basic flaw of the concept of external validity, namely, that it is based on an inductivist philosophy of science.

The reservation even Campbell and Stanley (1966) themselves appeared to harbor with regard to the concept of external validity did not prevent others from forging into the unknown, suggesting strategies that in their opinion would solve the problem. A first suggestion that has been proposed is to use samples that are (more) representative of the (world) population. For example, Arnett (2008) argued in an article published in *American Psychologist* that "research on the whole of humanity is necessary for creating a science that truly represents the whole of humanity" (p. 603). Similarly, Aronson, Wilson, and Akert (2005) suggested in their widely used textbook that "the only way to be certain that the results of an experiment represent the behavior of a particular population is to ensure that the participants are randomly selected from that population" (p. 45). They admitted that this approach would not be practical but added that "concerns about practicality and expense are not good excuses for doing poor science" (p. 46).

Apart from being unpractical, the concept of a representative sample makes sense only if a certain population is finite (e.g., the current population of the United States). If psychologists want to establish laws that apply to all of humanity, which they may typically aim for, they should be aware of the fact that "all of humanity" is not a finite population, because is also includes past and future generations. Furthermore, as suggested by Stroebe and Nijstad (2009), in a critical response to Arnett (2008), average results found with such "representative" samples might not apply to most members of that population. Suppose a researcher was interested in the impact of one-sided versus two-sided communications without realizing that the effect is moderated by intelligence (i.e., one-sided communications have more impact with less intelligent people, two-sided communications have more impact with more intelligent individuals; Hovland, Lumsdaine & Sheffield, 1949). If that researcher did the study with a representative sample of the U.S. population, the result would be a null effect, a finding that would not be true for most members of the population studied. Thus, in the absence of good theory about relevant population characteristics, even representative samples do not necessarily lead to more valid conclusions about populations than convenience samples.

As a second—and less cumbersome—solution, Henrich et al. (2010) suggested that universities should invest in creating "non-student subject pools—for example, by setting up permanent psychological and behavioral testing facilities in bus terminals, Fijian villages, rail stations, airports, and anywhere diverse where subjects might find themselves with extra time" (p. 82). In other words, replace subject pools representative of students, with pools representative of bus, rail and airplane travelers, or at least those subsection of travelers destined for stations or airports near the university.

For social psychologists, the use of such diverse samples would also cause problems for the operationalization of their independent variables, because

manipulations are assumed to create the same social condition for all participants. For example, in a study of attitude change, one has to use messages that argue a position with which most subjects are likely to disagree; otherwise there can be no attitude change. If such a study were conducted at a bus terminal, train station, or airport, it would be difficult to predict the original attitudinal position of the potential respondents. When social psychologists operationalize a theoretical variable, they need a homogenous population to be certain that their manipulation reflects the same variable in all their subjects. This is already problematic when studies are replicated with student participants of a different nationality or of the same nationality a decade later (Stroebe & Strack, 2014). But the task would be impossible with randomly chosen travelers or a random sample of a population (see also Baumard & Sperber, 2010, for a similar critique of cross-cultural studies).

Instead of attempting to increase external validity by increasing the diversity of participant samples, Simons et al. (2017) suggested a third, but equally problematic, solution. They proposed that empirical manuscripts should include a statement of the Constraints on Generality (COG) as a declaration that "explicitly identifies and justifies the target populations for the reported findings." This COG statement is meant to justify all claims of generality for the participants of the study, the stimulus materials, the experimental procedures, and the historical context in which the study was conducted. Specifically, the representativeness of these elements for the populations to which the findings are being generalized should be made explicit, on the basis of either empirical data or theoretical predictions. For example, researchers have to explain if the sample of experimental participants is representative for "all psychology undergraduates, all undergraduates in the United States, all adults, or even all mammals." According to the authors, the inclusion of COG statements will decrease the chance that the author's claim will be proven to be "embarrassingly" more limited than originally implied, it will increase the likelihood that subsequent replications by other researchers will be successful, and it will inspire follow-up studies.

Quite apart from the fact that, as we argue next, such statements are superfluous in theory-testing research, they would also be impossible to make. Suppose we demonstrated a particular finding among students at the University of Groningen. How would we know whether students at the University of Groningen are representative of all Dutch students? This issue has never been studied empirically, and given the weather conditions in Groningen, it is quite possible that students there differ in some way from, say, students in the southerly town of Maastricht, which enjoys a more clement climate. And Dutch psychology students are likely to differ from German students, who must have an excellent grade point average to be admitted to psychology, whereas there is no such selection in the Netherlands. But even if we demonstrated that our results could be replicated with student samples all over the Netherlands and Germany, the question would arise whether this causal effect would also hold for less educated or for older people or for other communicators, other communications, other attitude issues, and other methods of measuring attitudes.

Our discussion illustrates the problems of induction. As Popper (1959, Chapter 1) argued with reference to Hume (1748), it is impossible to derive general laws from singular observations. The problem cannot be solved by using representative samples of the world population or, on a smaller scale, travelers in bus stations, train stations, or airports. Fortunately, as we argue next, hypothetico-deductivism provides a workable solution to this problem.

When is representativeness of samples important?

Although representativeness of samples is not important for theory-testing research, it can be important for many forms of applied research. For example, in survey research, where researchers attempt to determine the percentage of members of a population that has a particular characteristic (e.g., holds a particular attitude, intends to vote for a particular candidate, or buy a particular product), representativeness of samples is important. Such questions do not concern laws of causal relations but simple facts: One wants to assess how certain features are distributed in a population that is finite but too large for a complete survey. For such questions, one needs a sample that is representative in the sense that the *relevant* feature is distributed in the sample approximately as it is in the population. Random sampling is one of the means for that purpose. We may speak here of *statistical representativeness.*

Representativeness is also important in experimental research that addresses certain applied issues. If researchers are interested in whether a particular treatment intervention is effective for a specific target population, they must test this intervention with respondents who are representative for this

target population. For example, in studies of consumer research one might be interested in the effectivity of a specific advertisement. It is therefore important in such studies that the impact of that advertisement is tested with the same type of individuals that form the target group for the advertisement. For this reason, the fact that Campbell and Stanley (1966) intended their analysis for education researchers is highly relevant for their conceptualization of external validity, a point also made by Mook (1983).

It is important to note, however, that the distinction between applied and theory-testing research does not refer to mutually exclusive categories, because applied research can also be theory testing. If a particular intervention has been derived from a theory, an intervention design can be used to assess the validity of that theory. For example, intervention studies were used to test a medical theory about the causes of obesity developed by the science journalist Gary Taubes (2011). He rejected the "energy in/energy out" model of obesity and suggested that obesity was caused not by an imbalance of calorie intake relative to energy output but by the consumption of the *wrong* calories, namely, carbohydrates rather than fat. Carbohydrates are the main driver of insulin secretion. If carbohydrate intake is restricted, insulin secretion falls and fat cells will release fatty acids resulting in weight loss. If this theory were valid, obese individuals put on a calorie-restricted low carbohydrate/high fat diet would be more likely to lose weight than obese individuals put on a high carbohydrate/low fat diet that contains the same calories (i.e., isocaloric). However, this theory was not supported in (applied) intervention studies (e.g., Foster et al., 2010; Naude et al., 2014). Thus, the decisive difference with regard to the need of representativeness of samples is not between theory testing and applied research but between research that tests theories and causal relations and research that does not (i.e., surveys and atheoretical intervention studies).

For the sake of argument, let us assume that Taubes's (2011) theory were valid for half of humanity but invalid for the other half. If half of humanity would gain weight on a diet rich in carbohydrates but low on fat, whereas for the other half a diet high in fat but low in carbohydrates would result in weight gain, then both types of diet would result in similar weight gain, if applied indiscriminately to—and averaged across—both types of human beings. However, what would be wrong here would not be the empirical study but the theory being tested. Taubes's assumption that his theory applied to all of humanity would be false and the theory would need to be refined as to which part of humanity it applied.

This is also a good example for demonstrating how the failure to support a theory could result in theoretical improvements. For example, if Taubes had conducted these studies himself, he might have used the empirical data to test for potential moderators. The elegant way would be to do this theory-guided. But less elegantly, he could also divide the samples into participants who lost weight with a given diet and those who gained weight with that diet. If such characteristics were found, one could incorporate them as moderators into the original theory. However, one would then have to conduct a *further* study to test this enriched theory with a new sample of participants.

The history of social psychology is full of examples of theories that have been modified in the light of inconsistent findings. For example, the theory of reasoned action (Fishbein & Ajzen, 1975) was developed in response to a devastating report by Wicker (1969) that attitudes were poor predictors of behavior. The elaboration likelihood model (Petty & Cacioppo, 1986) was developed to explain various inconsistencies in the attitude change literature that seemed to be related to differences in recipient involvement or issue knowledge. Finally, the incorporation of negativity of consequences (Cooper & Worchel, 1970) and freedom of choice (Linder, Cooper, & Jones, 1967) into dissonance theoretical explanations of the consequences of insufficient reward for counterattitudinal behavior was motivated by repeated failures to replicate the original Festinger and Carlsmith (1959) findings. In each case, the amended theories were then tested in new experimental studies that manipulated the assumed moderators. It is important to note, however, that none of these inconsistencies was discovered while replicating studies in airports, bus terminals, or Fijian villages.

Beyond representativeness and external validity: The hypothetico-deductive approach

Theories consist of a number of abstract concepts that reflect theoretical constructs and of hypotheses about the relationship between these constructs. If one tests predictions derived from a general theory, there is no problem of generalization and external validity, because the theory defines the population to which it applies. With psychological theories the assumption is usually that they apply to all human beings. This assumption is typically implicit and can be inferred from the fact that the theory does not specify a

particular subgroup of humanity to which it applies. For example, the frustration-aggression hypothesis of Dollard, Doob, Miller, Mowrer, and Sears (1939) assumed that frustration was the sole antecedent of aggression in all human beings. In testing that theory, one examines whether a causal relationship exists between the two theoretical variables frustration and aggression.

The abstract concepts that constitute a psychological theory are unobservable variables. To be able to test the theory empirically, these unobservable theoretical concepts have to be operationalized, that is, translated into observable terms in empirical hypotheses. In experiments designed to test whether frustration leads to aggression, frustration might be operationalized by having participants receive a negative evaluation from another person (confederate) on a task they had just performed, and aggression may be measured by giving the frustrated participant the opportunity to deliver some noxious stimulus to the frustrator (e.g., an electric shock). These assumptions that link unobservable theoretical concepts to empirical manipulations or measures are "auxiliary hypotheses" (Gadenne, 1984; Trafimow, 2012) that can themselves be true or false.

Unlike education researchers who in their applied research are interested in whether a particular type of training will improve a particular type of learning, researchers testing the frustration-aggression hypothesis are not interested in the determinants of the delivery of electric shocks. They are interested in whether frustration results in aggression, and they assume that their operationalization of the independent variable actually reflects the theoretical concept (i.e., frustration) they were trying to manipulate and that the dependent variable measured the concept (i.e., aggression) they were attempting to assess. Thus, the *validity* of theory-testing research depends not only on the internal validity of the experimental procedures but also on the validity of the experimenters' auxiliary hypotheses (Trafimow, 2012) that guided them in developing their operationalizations.

If experimental subjects, who have been frustrated, deliver more (or more severe) electric shocks to the frustrator than subjects who were not frustrated, the hypothesis is confirmed. There will be no question of generalization, because the theory specified the class of people to whom it applies, namely, "all of humanity." Thus, the method of theory testing solves the problem of generalization. "We can gain general knowledge by testing theories. Such knowledge consists of the theories that have been confirmed" (Gadenne, 2013, p. 6).

This does not mean, however, that such confirmation *proves* a theory to be true. As Popper (1959) argued, a theory can never be proven true. One reason for this is that there is always the possibility that researchers' auxiliary hypotheses were invalid or that, due to deficits in experimental control, third variables were responsible for the observed relationship between the variables manipulated or measured in a study. Nevertheless, a theory can be more or less well-supported depending on the number of strict empirical tests the theory has successfully undergone (Gadenne, 1984, 2013; Popper, 1959).

For the same reason, however, failure of a single experiment to support a hypothesis does not falsify a theory. If experimental subjects, who have been frustrated, fail to deliver more (or more severe) electric shocks to the frustrator than subjects who were not frustrated, one can always argue that the frustration manipulation failed to induce frustration or that the delivery of electric shocks was not a good measure of aggression. Such criticism has repeatedly been raised against most of experimental aggression research (e.g., Ritter & Eslea, 2005; Tedeschi & Quigley, 1996). However, even though the failure of one empirical study to support a theoretical prediction does not falsify a theory, repeated failure does raise serious doubts. One possibility of salvaging the theory is to more clearly specify the conditions under which a theoretical prediction would be supported (Trafimow, 2009). Examples for this in social psychology have been described earlier (e.g., Cooper & Worchel, 1970; Linder, Cooper, & Jones, 1967). Theories are abandoned only if better theories are developed that have higher empirical content. Such theories need to explain all the findings of the ones they replace (including those that are inconsistent with that theory) but also make additional predictions that could not be derived from the original theory.

In contrast to this hypothetico-deductive approach, the criticism of Henrich et al. (2010) seems to be based on an inductivist theory of science that assume that we generalize from the empirical findings to reality rather than deriving interpretations of reality from our theories. In a section criticizing that researchers often assume that their findings are universal, they argue, "Sampling from a thin slice of humanity would be less problematic if researchers confined their interpretations to the populations from which they sample" (p. 63). As we pointed out in the previous section, in theory-testing research such a restriction is

unnecessary: The aim is not to generalize but to test whether an effect that is predicted by the theory does actually occur. This is fortunate, because if the findings of our research would apply only to college students, the science of psychology would be useless for explaining the behavior of most of humanity and applications of psychological science to consumer, health, or economic behavior would be futile.

Can diversity be useful in theory testing?

We argued earlier that conducting theory-testing research in bus terminals, train stations, or airports or on representative samples of the world population is not very informative. If we find that our experimental results are the same with bus travelers as with university students, we would still not know whether our findings would also generalize to travelers passing through the local airport or through a different bus terminal. And if the experiment worked with university students but not with bus travelers or passengers passing through the local airport, one would not know why this was the case. Was it because travelers are older or less educated than university students? Or was it that some travelers did not understand the experimental instructions? Or perhaps that they were under stress, because they did not want to miss their bus or flight and were therefore less involved in the experiment than university students, who participated in their free time?

And yet, replicating experimental findings with varied samples can be useful if the selection of these samples is *theory guided*. After all, the assumption that a theory applies to all human beings is as much a *testable hypothesis* as the assumption that the auxiliary hypotheses that guided the operationalizations of theoretical constructs were valid. The important point here is that it is irrelevant whether aspects of the experimental situation are representative of "reality" (i.e., external validity) but whether they are valid empirical translations of the corresponding terms in the theoretical proposition under investigation. And as this correspondence assumption can be questioned with regard to the auxiliary hypotheses guiding the operationalization of theoretical variables, one can also ask whether college students are representative of humankind.

However, any challenge to the correspondence assumption would have to be *theory specific*. Thus, as one would ask in a test of the frustration-aggression hypothesis whether delivering electric shocks to another person is an optimal operationalization of *aggressive* behavior (i.e., a good representation of this theoretical construct), one could ask whether college students are representative of humankind *with regard to the theoretical hypothesis being tested*. This last aspect—that the difference has to be related to the research question being tested—is important, because it does not matter that college students are WEIRD, as long as their WEIRDness does not affect the theoretically predicted relation that is being tested.

Is cross-cultural research the answer?

Cross-cultural research would seem to provide the appropriate method to assuage doubts about the assumed universality of our psychological laws. However, like research involving travelers in bus terminals or airports, cross-cultural research is rarely informative unless researchers have specific hypotheses why the relationship between the theoretical concepts should differ between cultures or nations. The blind repetition of theory-testing research in different countries or different cultures is a futile exercise. If an experimental finding were replicated in some countries/cultures, but not in others, it would be difficult to know whether the problem was due to the invalidity of the universality assumption or to a failure to create the same experimental conditions across the different countries/cultures.

This would be less of a problem with simple perception studies such people's susceptibility to the well-known Mueller-Lyer illusion (i.e., participants who are shown a stylized arrow and are asked to place a mark on the midpoint of the figure, tend to place it more toward the tail end). When Rivers (1901) demonstrated that Murray islanders were less susceptible to the illusion than Europeans, he explained the difference with the assumption that Europeans live in more "carpentered" environments characterized by straight lines, right angles, and square corners than Murray islanders. This conclusion was later challenged by G. Jahoda (1971), who compared members of an African tribe, who lived either in a traditional rural environment or in African cities. He found no meaningful difference between these two groups and suggested that differences in retinal pigmentation between Europeans and dark-skinned people could be responsible for the differences in their susceptibility to the optical illusion. However, this interpretation was later refuted by Berry (1968), who compared Samples of Eskimos of Baffin Island and Temne of Sierra Leone, who lived in either a traditional or a moderately carpentered environment, and found for both groups that they showed more susceptibility to the Mueller-Lyer illusion if they lived in a more carpentered

environment. This research demonstrates that even with such simple stimuli and a relatively clear theoretical interpretation, testing this explanation cross-culturally is a complex task.

But as shown by the "cross-cultural experiments on threat and rejection," conducted by the "Organization for Comparative Social Research" and directed by Stanley Schachter as research coordinator, findings of cross-cultural research become uninterpretable with complex social situations and in the absence of clear theoretical predictions (Schachter et al., 1954). This group of researchers wanted to examine whether the original findings on rejection of deviates in group settings reported by Schachter (1951) could be replicated in seven countries, namely, Netherlands, Sweden, France, Norway, Belgium, Germany, and England. Even though conditions for this research were optimal, the result was a resounding failure. Manipulation checks showed that some or all of the experimental manipulations were unsuccessful in three of the seven locations (England, Belgium, and Germany). But even in the countries in which the manipulations appeared to have worked, their impact on the sociometric ratings differed widely: The Norwegian results were distinctly different from those of the Dutch, Swedish, and French experiments. Schachter et al. (1954) suggested a number of alternative interpretation for this inconsistency: "(a) The Norwegian results are attributable to experimental artifacts. (b) There are cultural differences between Norway and the other three countries. (c) The relationship among our variables is more complex than we originally thought" (p. 429).

Given these rather discouraging findings, it is not surprising that this was the only cross-cultural experiment attempted by the Organization for Comparative Social Research. Yet there is an important lesson to be learnt from this failure. As Roberts (1970) concluded in a review of cross-cultural studies:

> This study points to an important problem. When experimental replications in foreign lands obtain the same results as the original work, there is little difficulty in interpretation. When the results differ, and the strategy cannot be replicated across national boundaries, there is little hope of interpretation, particularly when we have no clear definition of culture. (p. 338)

In the meantime, mainly due to the dimensional approach to culture first promoted by Hofstede (1980), we have information about differences between cultures that are likely to be of relevance for individual behavior. Hofstede originally identified four dimension of culture— *individualism-collectivism*, *power distance*, *uncertainty avoidance*, and cultural *masculinity-femininity*—that provide an organizing structure that allows one to describe cultures. Of these dimensions, the *individualism-collectivism* dimension has had the strongest impact on research on cultural psychology (Smith, 2010), especially after researchers elaborated the variation in psychological processes along this dimension. Compared to members of individualistic countries, members of collectivist countries hold different construals of self, others, and the interdependence of the two (Markus & Kitayama, 1991). Members of individualistic countries such as the United States, the United Kingdom, or the Netherlands think of themselves as autonomous individuals. They tend to prefer *independent* relationships with others and to subordinate the goals of their in-groups to their own personal goals. In collectivistic countries such as China or Korea, people tend to think of themselves as members of their groups and to prefer *interdependent* relationships to others and to subordinate their personal goals to those of their in-groups. Related to these cultural variations in social orientation are variations in *cognitive style* along the analytic-holistic dimension. Nisbett, Peng, Choi, and Norenzayan (2001) proposed that members of Western nations process information in an analytic way, identifying the key elements in a situation and ascribing causality to focal actors. In contrast, members of Asian countries process information in a holistic way, paying attention to the total perceptual field and the relationship between all the elements in that field.

These differences are important because the countries in which most psychological research is being conducted (United States, United Kingdom, the Netherlands, and Germany) are among the most individualistic countries (Hofstede, 1980). They also allow us to identify certain areas of psychological research where experimental findings might differ between individualist and collectivist cultures. A great deal of cross-cultural research has confirmed these expectations (Kitayama & Cohen, 2007). The important difference between this new type of cross-cultural research and studies like the one conducted by Schachter et al. (1954) is that the former is theory guided, whereas the latter was not. We now have information about cultural differences that allow us to derive hypotheses from psychological theories about potential differences in findings of studies conducted in these different cultures. Because a general review of cultural psychology is beyond to scope of this article (we refer the interested reader to Kitayama & Cohen, 2007), we discuss cultural differences with regard to two paradigms to illustrate this approach.

The Asch (1951) conformity paradigm is an experimental situation that should be affected by the differences in social orientation between members of individualistic and collectivist countries. The paradigm assesses the influence of a unanimous majority on an individual group member in a situation in which the majority gives a false judgment. Because members of collectivist countries prefer *interdependent* relationships to others and are assumed to subordinate their personal goals to those of their in-groups, one would expect them to show more conformity than members of individualistic countries. Indeed, in a meta-analysis of 133 experiments, Bond and Smith (1996) found that the rate of conformity was lower in individualistic than collectivistic countries.

In view of the differences in cognitive style of members of individualistic and collectivistic countries, the correspondence bias (also referred to as the fundamental attribution error) is a phenomenon where one would expect to observe cultural differences. The correspondence bias refers to a tendency to ascribe to a person an attitude that corresponds to that person's behavior, even when the behavior was clearly determined by the situation. Because members of collectivist nations are assumed to process information in a holistic way, paying attention to the total perceptual field and the relationship between all elements in that field, one would assume that they are less likely to be subject to the correspondence bias. This hypothesis was supported by Choi and Nisbett (1998) in a study with American and Korean subjects.

What can we conclude from these studies with regard to the universality assumption of psychological theories? Obviously, both the levels of conformity observed in the Asch situation and the correspondence bias varied across the countries or cultures studied. To scientists taking an inductivist position, these cultural differences would be a clear indication that psychological laws supported by studies with college students do not apply to members of these other cultures. In contrast, scientists taking a Popperian (1959) position would not find these cultural differences problematic. The important question from a hypothetico-deductive perspective is not whether a phenomenon is invariant across different cultures (i.e., a cultural universal) but whether these differences can be explained by our psychological theories. The fact that the results of these cross-cultural studies supported the hypotheses of the researchers who conducted these studies clearly suggests that these cross-cultural differences could be predicted from psychological theories that were originally developed in an individualistic culture and tested with American undergraduate students.

Conclusions

The fact that one of the most influential methodological contributions to psychology was originally written for a handbook of research on education may have been responsible for a widespread confusion about generalizability of research findings. The concept of external validity introduced by Campbell and Stanley (1966) for applied research in education was interpreted as implying that general laws can be derived from empirical observations. If that were correct, the fact that most psychological research is conducted with samples of college students (or more recently MTurk workers) would indeed call the whole of scientific psychological knowledge into question.

In this article, we reminded our readers that this type of inductivism has been rejected since Hume (1748): Empirical findings among any type of population can never be generalized to other populations. Instead, empirical observations are used to test hypotheses derived from theories, and it is the theories that specify the domain of applicability. Most psychological theories assume that this domain includes the whole of humanity. However, even though theories can vary in the extent to which they are empirically supported, they can never be proven to be true: Any aspect of a theory can be challenged. We can question whether the assumptions made by a theory about the causal relationship between the various theoretical states or events is correct; we can doubt that the auxiliary hypothesis that link the theoretical concepts to experimental manipulation and/or measures are valid; finally, we can also question the universality assumption that our theory applies to all of humanity. We can point out that experimental subjects with whom a theory has been tested differ from the rest of humanity in psychological characteristics that are relevant to the theory in question. Whereas it is not informative to blindly replicate studies with people passing through bus terminals, train stations, or airports, or among representatives of different cultures, theory-guided research with subject populations that differ from the original research participants in psychological characteristics that are relevant for the theory being tested have resulted in important findings.

Some of this research has indicated that phenomena (e.g., the fundamental attribution bias) assumed to be cultural universals do in fact vary across selected cultures or that the extent to which individuals

conform to unanimous majorities increases with the collectivist nature of the culture of which the research participants are members. This is problematic for inductivists, because it indicates that one cannot generalize from empirical observations of how people behave or think in one culture, to thoughts or behaviors of members of another culture. As we have argued, this cultural variance does not pose a problem for a hypothetico-deductive approach to scientific research. The important question from a hypothetico-deductive perspective is not whether psychological phenomena are culturally invariant but whether the observed variance can be explained by our psychological theories. This was definitely the case in the research examples reviewed earlier.

We would like to avoid the inductivist fallacy to generalize from our limited knowledge of cross-cultural research to all of that research or even to research that will at some point be conducted in the future. Thus, we can only point out that there is great deal of cross-cultural evidence that our psychological laws are valid beyond the student (or more recently MTurk) populations with whom they have been validated. More important for the present argument, however, is that diversifying our subject populations is going to bring an advance in psychological knowledge only when we have good theories on why relations among variables may be different in different populations.

Notes

1. The "replication crisis" refers to the fact that empirical findings in psychology often do not fare so well in direct replication studies (e.g., Pashler & Wagenmakers, 2012; Stroebe, 2016; Stroebe & Strack, 2014).

2. For pragmatic reasons and brevity, we use the term "law" or "psychological law" throughout the article. However, typically findings in psychology are more restricted in scope and lack the robustness of the "laws" in, for example, the natural sciences.

3. Popper's methodology is controversially discussed in Lakatos and Musgrave (1970). For a reformulation and defense of Popper's view, see Andersson (1994) and Musgrave (1999).

References

Andersson, G. (1994). *Criticism and the history of science: Kuhn's, Lakatos's and Feyerabend's criticisms of critical rationalism*. Leiden, Netherlands: Brill.

Arnett, J. (2008). The neglected 95%: Why American Psychology needs to become less American. *The American Psychologist*, *63*, 602–614. doi:10.1037/0003-066X.63.7.602

Aronson, E., Wilson, T. D., & Akert, R. M. (2005). *Social Psychology* (5th ed.). Saddle River, NJ: Prentice Hall.

Asch, S. E. (1951). Effects of group pressure upon modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177–190). Pittsburgh, PA: Carnegie.

Baumard, N., & Sperber, D. (2010). Weird people, yes, but also weird experiments. *Behavioral Brain Sciences*, *33*, 84–85. doi:10.1017/S0140525X10000038

Berry, J. W. (1968). Ecology, perceptual development and the Müller-Lyer illusion. *British Journal of Psychology*, *59*, 205–210. doi:10.1111/j.2044-8295.1968.tb01134.x

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's Line Judgment Task. *Psychological Bulletin*, *119*, 111–137.

Calder, B. J., Phillips, L. W., & Thybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research*, *9*, 240–244.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297–312. doi:10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Choi, I., & Nisbett, R. E. (1998). Situational salience and cultural differences in the correspondence bias and actor observer bias. *Personality and Social Psychology Bulletin*, *24*, 949–960. doi:10.1177/0146167298249003

Christie, R. (1965). Some implications of research trends in social psychology. In O. Klineberg & R. Christie (Eds.), *Perspectives in social psychology* (pp. 141–152). New York: Holt, Rinehart & Winston.

Cook, T. D., & Campbell, D. T. (1979). *Quasi experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cooper, J., & Worchel, S. (1970). Role of undesirable consequences in arousing cognitive dissonance. *Journal of Personality and Social Psychology*, *16*, 199–206. doi:10.1037/h0029830

Dollard, J., Doob, L. W., Miller, N. E., Mowrer, O. H., & Sears, R. R. (1939). *Frustration and aggression*. New Haven, CT: Yale University Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal Psychology*, *58*, 203–210. doi:10.1037/h0041593

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Foster, G. D., Wyatt, H. R., Hill, J. O., Makris, A. P., Rosenbaum, D. L., Brill, C., … Klein, S. (2010). Weight and metabolic outcomes after 2 years on a low-carbohydrate versus low-fat diet: A randomized trial. *Annals of Internal Medicine*, *153*, 147–157. doi:10.7326/0003-4819-153-3-201008030-00005

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3–12. doi:10.1177/1088868313507536

Gadenne, V. (1976). *Die Gültigkeit psychologischer Untersuchungen. (The validity of psychological research)*. Stuttgart, Germany: Kohlhammer.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung. (Theory and experience in psychological research)*. Tuebingen: Mohr und Siebeck.

Gadenne, V. (2013). External validity and the new inductivism in experimental economics. *Rationality, Markets and Morals*, *4*, 1–19. http://www.rmm-journal.de/.

Gage, N. L. (1963). *Handbook of research on teaching*. Chicago: Rand McNally.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–135. doi:10.1017/S0140525X0999152X

Hovland, C. J., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments on mass communication*. Princeton, NJ: Princeton University Press.

Hofstede, G. H. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Hume, D. (1748). *An enquiry concerning human understanding*. New York: Bobbs-Merrill.

Jahoda, G. (1971). Retinal pigmentation, illusion susceptibility and space perception. *International Journal of Psychology*, *6*, 199–208. doi:10.1080/00207597108246683

Kitayama, S. (2017). Editorial. Journal of Personality and Social Psychology: Attitudes and social cognition. *Journal of Personality and Social Psychology*, *112*, 357–360.

Kitayama, S., & Cohen, D. (2007). *Handbook of cultural psychology*. New York: Guilford.

Kruglanski, A. W., & Kroy, M. (1973). Outcome validity in experimental research: A reconceptualization. *Journal of Representative Research in Social Psychology*, *7*, 168–178.

Linder, D. E., Cooper, J., & Jones, E. E. (1967). Decision freedom as a determinant of the role of incentive magnitude in attitude change. *Journal of Personality and Social Psychology*, *6*, 245–254. doi:10.1037/h0021220

Lakatos, I., & Musgrave, A. (1970). *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.

Markus, H., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*, 289–374.

Mook, D. (1983). In defense of external validity. *American Psychologist*, *38*, 379–387.

Musgrave, A. (1999). *Essays on realism and rationalism*. Amsterdam: Rodopi.

Naude, C. E., Schoonees, A., Senekal, M. et al. (2014). Low carbohydrate versus isoenergetic balanced diets for reducing weight and cardiovascular risk: A systematic review and meta-analysis. *PLoS One*, *9*, e100652.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, *108*, 291–310.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in the social sciences: A crisis of confidence. *Perspectives on Psychological Science*, *7*, 528–530.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). New York: Academic Press.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Ritter, D., & Eslea, M. (2005). Hot sauce, toy guns, and graffiti: A critical account of current laboratory aggression paradigms. *Aggressive Behavior*, *31*, 407–419. doi:10.1002/ab.20066

Rivers, W. H. R. (1901). Introduction and vision. In A. C. Haddon (Ed.), *Reports of the Cambridge anthropological expedition to the Torres Straits* (Vol. *2*, pp. 117–128). Cambridge, England: Cambridge University Press.

Roberts, K. H. (1970). Looking at an elephant: Evaluation of cross-cultural research related to organizations. *Psychological Bulletin*, *47*, 327–350.

Schachter, S. (1951). Deviation, rejection, and communication. *Journal of Abnormal Psychology*, *46*, 190–206. doi:10.1037/h0062326

Schachter, S., Nuttin, J., de Monchaux, C., Maucorps, P. H., Osmer, D., Duijker, H., … Israel, J. (1954). Cross-cultural experiments on threat and rejection. *Human Relations*, *7*, 403–439. doi:10.1177/001872675400700401

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515–530. doi:10.1037/0022-3514.51.3.515

Simons, D., Yuichi, S., & Lindsay, D. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. Retrieved from https://osf.io/preprints/psyarxiv/w9e3r

Smart, R. (1966). Subject selection bias in psychological research. *Canadian Psychologist*, *7a*, 115–121. doi:10.1037/h0083096

Stroebe, W. (2016). Are most published social psychological findings false? *Journal of Experimental Social Psychology*, *66*, 134–144. doi:10.1016/j.jesp.2015.09.017

Stroebe, W., & Nijstad, B. A. (2009). Do our psychological laws apply only to Americans? *American Psychologist*, *64*, 569. doi:10.1037/a0016090

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71. doi:10.1177/1745691613514450

Taubes, G. (2011). *Why we get fat*. New York: Anchor Books.

Tedeschi, J., & Quigley, B. (1996). Limitations of laboratory paradigms for studying aggression. *Aggression and Violent Behavior*, *1*, 163–177. doi:10.1016/1359-1789(95)00014-3

Trafimow, D. (2009). The theory of reasoned action: A case study of falsification in psychology. *Theory & Psychology*, *19*, 501–518. doi:10.1177/0959354309336319

Trafimow, D. (2012). The role of auxiliary assumptions for the validity of manipulations and measures. *Theory & Psychology*, *22*, 486–498. doi:10.1177/0959354311429996

Wicker, A. W. (1969). Attitude versus action: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*, 41–78. doi:10.1111/j.1540-4560.1969.tb00619.x

Wintre, M. G., North, C., & Sugar, L. A. (2001). Psychologists' response to criticism about research based on undergraduate participants: A developmental perspective. *Canadian Psychology*, *42*, 216–225. doi:10.1037/h0086893